GEOBIT: A Geodesic-Based Binary Descriptor Invariant to Non-Rigid Deformations for RGB-D Images

Erickson R. Nascimento[1], Guilherme Potje[1], Renato Martins[1,2], Felipe Cadar[1],
Mario F. M. Campos[1], and Ruzena Bajcsy[3]
[1]Universidade Federal de Minas Gerais (UFMG) [2]INRIA [3]University of California Berkeley
{erickson,guipotje,renato.martins,cadar,mario}@dcc.ufmg.br, bajcsy@eecs.berkeley.edu

Abstract

At the core of most three-dimensional alignment and tracking tasks resides the critical problem of point correspondence. In this context, the design of descriptors that efficiently and uniquely identifies keypoints, to be matched, is of central importance. Numerous descriptors have been developed for dealing with affine/perspective warps, but few can also handle non-rigid deformations. In this paper, we introduce a novel binary RGB-D descriptor invariant to isometric deformations. Our method uses geodesic isocurves on smooth textured manifolds. It combines appearance and geometric information from RGB-D images to tackle non-rigid transformations. We used our descriptor to track multiple textured depth maps and demonstrate that it produces reliable feature descriptors even in the presence of strong non-rigid deformations and depth noise. The experiments show that our descriptor outperforms different state-of-the-art descriptors in both precision-recall and recognition rate metrics. We also provide to the community a new dataset composed of annotated RGB-D images of different objects (shirts, cloths, paintings, bags), subjected to strong non-rigid deformations, to evaluate point correspondence algorithms.

1. Introduction

The ability to make sense of real-world objects from images, considering all possible variations of their characteristics such as texture, shape and deformation, is central for an adequate interpretation of scenes and objects in the world around us. The appearance of these objects on images is susceptible to a large number of conditions and transformations. For instance, while visually recognizing or tracking objects, we need to deal with partial view occlusions, rotations, and illumination changes, but also with the challenging condition of non-rigid surface deformations. Therefore, finding properties that characterize an object and remain invariant under such conditions, play a key role in the development of image recognition, tracking, and multiple view reconstruction. A common approach, to overcome the influence of these conditions, is to represent objects as a sampling of interest points, which are encoded with feature vectors called descriptors that contain distinguished features to characterize each object ubiquitously.

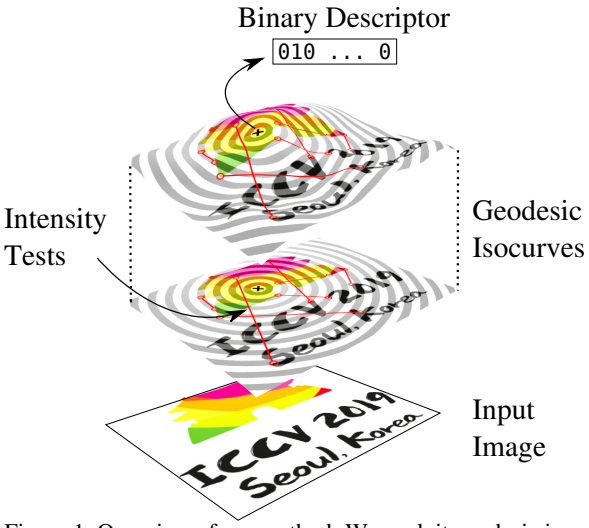


Figure 1. Overview of our method. We exploit geodesic isocurves of a textured 2D manifold subjected to isometric deformations.

In this paper, we introduce a new binary descriptor that combines appearance and geometric information from RGB-D images to handle isometric non-rigid deformations. Our method is invariant to image scale and uses geodesic isocurves on smooth textured manifolds. We used our descriptor to track multiple textured depth maps and demonstrate that it is robust and provides reliable results even in the presence of strong non-rigid deformations and depth noise. Figure 1 illustrates our descriptor.

Over the past few decades, numerous methodologies to extract features have been proposed (*e.g.*, [20, 6, 14, 13, 30, 22, 2, 21, 28]). These approaches can be roughly grouped

based on the type of input information, such as intensity or depth images. Even though image-based methods tend to exploit much of the rich information engrafted in images wisely, these techniques are restricted to 2D data. Thus, the performance of texture-based descriptors tends to quickly degrade with the decreased availability of texture and illumination in the scene. On the other hand, depth images have become increasingly used to define feature descriptors. Their information is less sensitive to lack of texture or illumination changes in the scene surfaces. Some examples of descriptors exploring the surface geometry are Spin-Image [6] and SHOT [26]. Despite the high discriminating power provided by these geometric descriptors, some issues still remain, such as their inherent algorithm complexity to compute the feature vector and, for some of them, the requirement of a large amount of data to avoid ambiguities.

Recently, leveraging both appearance (intensity images) and shape (from depth information) cues has been successfully adopted by many recent works to increase object detection recognition rate [7, 30] and matching [19, 17, 30, 27], boosted by the advent of low-cost RGB-D devices.

However, the majority of these methodologies, as mentioned earlier, are capable of detecting and extracting features only in the presence of a restricted number of transformations, such as rotations, scale, and translations [16]. For instance, an object can be deformed, *i.e*., the same object may assume different forms, which implies that other types of transformations are worth considering. Thus, unlike most methods, our descriptor takes a step towards using both visual and geometrical features to extract intrinsic properties to characterize real-world objects. In the experiments, our descriptor presented the highest point match scores, which in turn tend to benefit many tasks in computer vision, *e.g*., SfM, object detection, image recognition, tracking (shown in the paper), to name a few.

The main contributions of this paper can be summarized as follows: i) A lightweight binary keypoint descriptor that leverages appearance and geometrical information to extract deformation-invariant features; ii) A new RGB-D dataset with annotated matches and composed of synthetic and real-world objects subjected to a variety of non-rigid deformations.

2. Related Work

Extracting descriptors from images usually provides rich information on the object features, while geometrical information, produced by 3D sensors, is less sensitive to lighting conditions. A representative approach on images is the SIFT [14] descriptor. It first extracts features using local gradients and then estimates a characteristic orientation of the keypoint's neighborhood to provide invariance to rotation. A recent approach that has become popular is the use of binary strings to assembly the feature vector (*e.g*., [2, 21, 12]), which is highly inspired by the idea of Local Binary Patterns (LBP) presented by Ojala *et al*. [20]. The main advantage of using binary strings, to represent feature vectors, is their small computational cost and reduced storage requirements.

One of the enduring grand challenges in shape analysis is to extract properties that preserve the intrinsic geometry of shapes. Geodesic distances are well known intrinsic properties as far as isometric transformations are concerned. Kokkinos *et al*. [9] built the intrinsic shape context (ISC) descriptor based on properties of the geodesic distance. The work of Shamai *et al*. [23] proposed and evaluated a new basis for geodesic distance representation, as well as how to efficiently approximate the distance. Despite advances achieved by these works, their technique is most suitable to 3D shapes only. In the same direction, Guan *et al*. [4] proposed BRISKS, a geodesic-aware BRISK descriptor to spherical images. BRISKS, however, is designed to tackle solely 2-sphere manifolds, different from our descriptor that considers more general image deformations.

A few studies have addressed resolving the problem of matching keypoints on deformable surfaces. A representative approach that faces this problem is the work of Moreno-Noguer [16, 24]. They proposed a new framework to use kernels based on diffusion geometry on 2D local patches, named DaLI descriptor. DaLI is designed to handle non-rigid image deformations and illumination changes. Despite remarkable advances in extracting features invariant to non-rigid image deformations, we show in our experiments that our approach outperforms DaLI in terms of recognition rate, precision-recall and computational effort.

The use of multiple cues, such as texture and geometric features, has become popular in the last few years to improve the matching quality as well as to increase the discrimination power of feature vectors. To increase the recognition rate, Kanezaki *et al*. [7] proposed the global descriptor VOSCH which combines depth and texture. Another descriptor that uses both depth and texture is the Mesh-HOG [30]. The authors used a texture extracted from 3D models to create scalar functions defined over a 2D manifold. Similarly, Tombari *et al*. proposed an extension of their shape only descriptor SHOT [26] that incorporates texture [27]. This extension, called CSHOT, has signatures composed of two concatenated histograms: One that contains the geometric features and another encoding the texture information. Similarly, Lai *et al.* [10, 11] proposed to use two well-known descriptors for each type of data: SIFT, for image and Spin-Image for geometry, and then concatenate both to compose the feature vector. Lightweight descriptors that are able to combine geometrical and texture information were also proposed by Nascimento *et al*. [19, 17, 18]. The authors presented EDVD descriptor [19], which builds a rotation invariant represen-

tation based on the direction of the normals using an extended Gaussian image followed by the application of the Fourier transform. The BRAND [17] descriptor encodes information as a binary string embedding geometric and texture cues, and presents rotation and scale invariance.

In this work, we take a similar approach to improve the quality on matching keypoints by using the depth data to estimate intrinsic surface properties. Our technique builds a descriptor which takes into account both sources of information to create a unique representation of a region, simultaneously considering texture and shape. We employed our descriptor to track objects and, as our experiments demonstrate, the proposed descriptor significantly improves tracking accuracy, precision, and its robustness to strong isometric deformations beyond different scales and rotations.

3. Methodology

Our descriptor exploits visual and geometrical information to encode deformation-invariant features into a binary vector. On the one hand, the use of texture information results in a highly discriminative descriptor. On the other hand, depth information allow us to define the binary tests invariant to non-rigid deformations and scale.

Our descriptor receives as input an RGB-D image $\mathcal{F} = \{\mathcal{I},\ \mathcal{D}\}$, composed of an image $\mathcal{I} \in [0,\ 1]^{m \times n}$ as pixel intensities and $\mathcal{D} \in \mathbb{R}_+^{m \times n}$ as depth information, and a list of l detected keypoints $\mathcal{K} \in \mathbb{R}^{l \times 2}$. For each pixel $\mathbf{p} \in \mathbb{P}^2$, $\mathcal{I}(\mathbf{p})$ provides the pixel intensity and $\mathcal{D}(\mathbf{p})$ the respective depth. We split the method into two primary steps: After extracting the intrinsic surface properties (*i.e.*, geodesic distance), we select a set of pairs of pixels to create a gradient field to extract the visual pattern. Since our descriptor is built considering the geodesic distance, it provides invariance to scale in image space, and isometric surface deformations.

3.1. Geodesic Approximation with Heat Flow

In this section, we describe how to compute the geodesic distance of any two points in a 2D manifold using a diffusion strategy, named heat flow and proposed by Crane *et al.* [3]. Although other strategies could be used (*e.g.*, the fast marching algorithm [25]), the heat flow approximation brings us several advantages such as pre-factoring for efficiency and the possibility of being applied to point clouds and polygonal meshes.

Let $\mathbf{u} \in \mathbb{R}^{|\mathcal{V}|}$ be a piecewise linear function on a 2D manifold, *i.e.*, a simplicial complex mesh $\mathcal{M}$ that comprises a collection of triangles and vertices $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, where each edge is shared by at most two triangles. For each vector on a triangle with unit normal $\mathbf{N}$ and face area A_f, $\mathbf{e}_i^1$ and $\mathbf{e}_i^2$ are the two edge vectors incident to the vertex i, and u_i is the value at the opposing vertex. We denote the function $\phi : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}_+$ as the geodesic distance

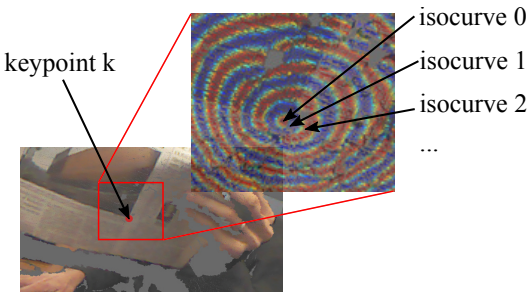


Figure 2. After approximating the geodesic distance using heat flow, we discrete the ϕ into isocurves of 4 cm size. Each test pair is localized using the isocurve id and the angle w.r.t. to the patch orientation.

approximation between any pair of vertices. In order to approximate the geodesic distance ϕ using the heat flow, we solve the Poisson equation:

$$\mathbf{L}_C \phi = \boldsymbol{\nabla} \cdot \mathbf{X}, \tag{1}$$

where $\mathbf{L}_C \in \mathbb{R}^{|V| \times |V|}$ is the cotangent Laplacian matrix, and $\boldsymbol{\nabla} \cdot \mathbf{X}$ contains the integrated divergences computed in the normalized vector field $\mathbf{X}$. In a 2D manifold sampled as a triangular mesh, the following divergence operator approximation holds:

$$\boldsymbol{\nabla} \cdot \mathbf{X} = \frac{1}{2} \sum \cot \theta_1 (\mathbf{e}_i^1 \cdot \mathbf{X}_j) + \cot \theta_2 (\mathbf{e}_i^2 \cdot \mathbf{X}_j), \tag{2}$$

where, for each vertex i, we sum over all adjacent triangles j of vertex i. The angles θ_1 and θ_2 are the opposing angles of vertex i and the vectors $\mathbf{X}_j$ are gathered from $\mathbf{X} = -\nabla \mathbf{u} / \|\nabla \mathbf{u}\|_2$, where the discrete gradient $\nabla \mathbf{u}$ can be computed as:

$$\nabla \mathbf{u} = \frac{1}{2A_f} \sum_i u_i (\mathbf{N} \times \mathbf{e}_i^1). \tag{3}$$

Finally, the $\mathbf{u}$ function using the heat flow, for a fixed time t, is given by solving the system $(\mathbf{A} - t\mathbf{L}_C)\mathbf{u} = \boldsymbol{\delta}_i$, where $\mathbf{A}$ is a diagonal matrix encoding the vertex areas and $\boldsymbol{\delta}_i$ is a vector with 1 in the i-th component and 0 in all others. We then define the set $\boldsymbol{\Phi}$ composed of isocurves after discretizing the ϕ function (see Figure 2). Since the geodesic distances are deformation-invariant properties, as far as isometric transforms are concerned, all pixels belonging to a specific isocurve will remain in the same isocurve after the surface deformation.

3.2. Binary Feature Extraction

After approximating the geodesic distance of the keypoint neighborhood, we can compute the visual features based on the binary gradient field around the keypoint. The idea behind this step is similar to the one used by the LBP [20], BRIEF [2], and more recently ORB [21].

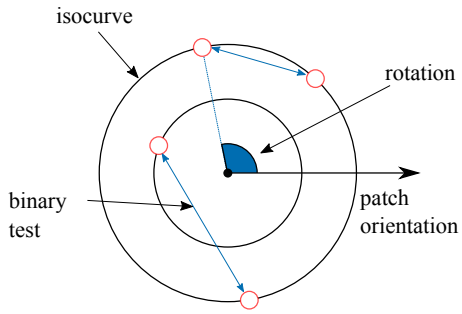


Figure 3. Example of two binary tests. We store for each binary test in the pattern the isocurve c and the rotation α w.r.t. the patch orientation of two points.

The gradient directions in this neighborhood are computed using image intensity difference tests, which have small memory requirements and processing time for matching. Given an image keypoint $\mathbf{k} \in \mathcal{K}$, assume an image patch $\mathbf{P}$ centered at $\mathbf{k}$. We sample pixel pairs around the keypoint $\mathbf{k}$ using a fixed pattern with locations given by a distribution (Figure 5 shows two tested distribution patterns). We store for each point in the pattern, the isocurve $c \in \Phi$ and the rotation α w.r.t. the patch orientation, as illustrated in Figure 3 with two test pairs of points lying onto two different isocurves. We can then build the set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$, as the fixed set of sampled pairs from $\mathbf{P}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ encode the isocurve and angle of the i-th pixel of the binary test pair, *e.g.*, $\mathbf{x_i} = (\alpha_i, c_i)^T$. Before constructing the visual feature descriptor, the patch $\mathbf{P}$ is translated to the origin and then rotated by the transformation $\mathbf{T}_\theta$, which produces a set

$$\mathcal{P} = \{(\mathbf{T}_\theta(\mathbf{x}_i), \mathbf{T}_\theta(\mathbf{y}_i)) | (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S}\}. \quad (4)$$

Thus, similar to DaLI, for each keypoint we compute a set of candidate descriptors (12 in our experiments) with different orientations by rotating the coordinates of the pattern points in set $\mathcal{S}$ using discretized rotations uniformly sampled from $[0, 2\pi]$, *i.e.*, adding $\theta = n\pi/6,\ n \in \{0, \ldots, 11\}$ to the first coordinate $\mathbf{T}_\theta(\mathbf{x}_i) = (\alpha_i + \theta, c_i)$. Then, in the matching step, we select the descriptor with orientation that results in the smallest distance between two compared descriptors. This strategy has shown better performance when compared to calculating the orientation for each keypoint using gradient-based approaches, mainly because non-rigid deformations around the keypoints introduce additional noise in the orientation estimation.

The extracted descriptor from the patch $\mathbf{P}$ associated with the keypoint $\mathbf{k}$ is then represented as the binary string:

$$b(\mathbf{k}) = \sum_{1}^{n} 2^{i-1} [p(\mathbf{x}_i) < p(\mathbf{y}_i)], \quad (5)$$

where $[t]$ is the Iverson bracket that returns 1 if the predicate t is true and 0 otherwise, and $p(\mathbf{x}_i)$ returns the corresponding pixel with $\mathbf{x}_i$ coordinates. The comparison in the bracket captures gradient changes in the keypoint neighborhood.

3.3. Sensitivity Analysis to Depth Errors and Computational Effort

The geodesic isocurve computation, described in Section 3.1, can be computationally intensive on high-resolution meshes. We argue that low-resolution depth images can be used to estimate the geodesic distance without degrading the results. The advantages of using lower resolution meshes are twofold. First, it dramatically increases the efficiency of the algorithm, since a smaller system of diffusion equations are solved, and a reduced number of operations are performed. Second, the diffusion operator is more robust to depth noise in the down-sampled smoothed mesh.

Therefore, we implement a multi-resolution strategy in three stages. First, we sub-sample the depth employing a Gaussian pyramid of depth two and used the lowest resolution depth image in the experiments. We employed an isometric bivariate Gaussian kernel of dimension five and unitary standard deviation. Then, we approximate the geodesic isocurves on the low-resolution mesh; finally, we up-sample the isocurves to the original resolution with bilinear interpolation. This multi-resolution strategy reduces the total number of vertices by a factor of $16\times$ and the algorithm runtime by at least a $35\times$ factor, and we show in the experiments that the smoothing can considerably increase the robustness to noise while keeping relevant geometric features.

4. Experiments

We evaluate the proposed method with both simulated and real data and compare the results for different descriptors. We adopt the recognition rate [2] and inverse of the precision-recall curve as metrics of comparison. We matched all pairs of keypoints from two images using brute force matching. Whenever the Hamming distance (BRAND, ORB, and ours) or Euclidean distance (DaLI and MeshHoG) is below a threshold, the pair is considered to be a valid match. We labeled valid matches with two keypoints corresponding to the same physical location (according to the ground-truth) as positive, and as negative otherwise. For the recognition rate metric, we consider the nearest neighbor of each descriptor in the other set as the predicted correspondence, which is used to calculate the accuracy rate.

4.1. RGB-D Non-Rigid Dataset

Real-world data. To evaluate the matching capability of our descriptor on real-world images, we built a new data set [1] composed of 6 deformable objects and a total of 74

[1] https://www.verlab.dcc.ufmg.br/descriptors/iccv2019

Figure 4. Examples of real-world and synthetic data in our dataset. The first two rows show examples of real data and the third row shows images of synthetic sequences.

pairs of RGB-D images captured with a Kinect™. All images were acquired at a resolution of 640×480 pixels. Different levels of isometric deformations were applied to each object. Naturally, non-linear illumination changes also arise when manipulating the surface of those objects. We manually annotated about 50 keypoints and the ground-truth correspondence for all datasets, since we cannot obtain a parametric model that describes arbitrary non-rigid deformations. The first two rows in Figure 4 show some examples of the real-world data in our dataset.

Synthetic data. We used a physic's simulation of cloth to create arbitrary non-rigid isometric deformations with ground-truth correspondences. In a nutshell, considering a grid of particles having mass and a 3D position, Newton's second law is applied in conjunction with Verlet integration, to act over the particles' position, *i.e.*, when forces like wind and gravity are applied. A constraint satisfaction optimization step is performed over all particles to enforce constant distance of neighboring particles, thus keeping the deformation isometric. The texture is applied onto the mesh generated by the grid and rendered with diffuse illumination as the cloth moves (which causes non-linear illumination changes). While the simulation is running, we uniformly sampled pixels from the image and used the Harris corner score to retain approximately 95 corner-like features. The synthetic data is composed of 18 pairs of images comprising three different textures with arbitrary deformations, and rotations. The third row in Figure 4 shows some examples of the synthetic data in our dataset.

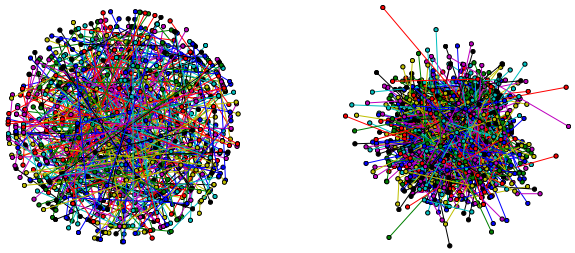


Figure 5. Binary tests patterns using uniform (on the left) and normal distributions (on the right). We tested both distributions and found that the Gaussian distribution results in slightly higher recognition rates.

4.2. Baselines and Metrics

We compared our results against the binary descriptor for 2D images ORB [21]; two descriptors that combines texture and shape: MeshHOG [30] and BRAND [17]; and the deformation-invariant descriptor DaLI [16].

Similar to Tombari *et al.* [2], we evaluate the matching performance using the recognition rate. Since we have annotated all corresponding keypoints for all pairs of images, we can compute reliably the number of corresponding keypoints between two images. We also evaluate the performance of our descriptor using precision-recall curves [8, 15]. Using a brute-force algorithm, we matched all pairs of keypoints from two different images. If the distance computed between descriptors dropped below a threshold t, the pair was considered a *valid match*. The *valid matches* are those for which two keypoints correspond to the same physical location (as determined by the annotation), and so defining the number of *true positives*. If the keypoints in a *valid match* come from different physical locations, then we increment the number of *false positives*. From these values, we compute the $recall$ and $1 - precision$. We report the area under the curve (AUC) of $recall$ vs. $1 - precision$ curves.

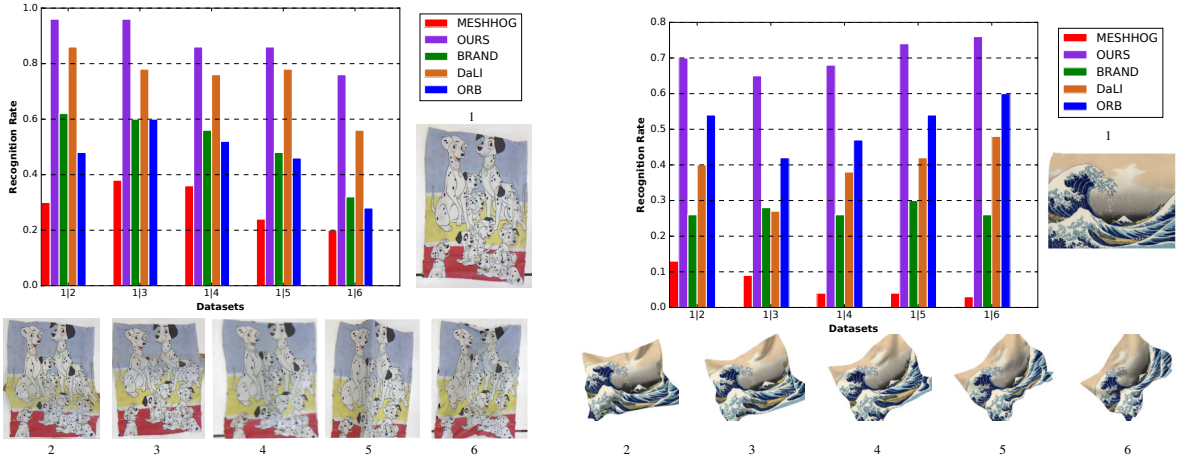
4.3. Parameter Settings

We empirically found the best values to be used as the angular isocurve size and the descriptor size. In this work, we set the isocurve size to 4 cm. We also tested different sizes of the feature vector, and we chose $1,024$ bits as the default size.

Binary tests distribution. Our descriptor performs binary tests in the neighborhood around the keypoint. This analysis is based on a set of pixels selected by a distribution function $\mathcal{S}$. We tested two different distributions. The pattern of each distribution is illustrated in Figure 5. Assuming that the origin of the patch coordinate system is located at the keypoint, we selected $1,024$ pairs of pixels using the following distributions: i) An isotropic Gaussian distribu-

Table 1. Comparison of our descriptor against standard methods. Our descriptor is able to provide higher recognition rate and AUC values.

Dataset (# pairs)	Recognition Rate					AUC				
	BRAND	DaLI	MeshHOG	ORB	Ours	BRAND	DaLI	MeshHOG	ORB	Ours
Shirt1 (14)	0.48	0.65	0.27	0.52	**0.73**	0.45	0.42	0.23	0.54	**0.75**
Shirt2 (18)	0.53	0.66	0.25	0.51	**0.74**	0.36	0.49	0.18	0.50	**0.54**
Shirt3 (17)	0.56	0.67	0.32	0.65	**0.72**	0.44	0.54	0.23	0.61	**0.63**
Can (6)	0.21	0.22	0.15	0.17	**0.23**	0.16	0.07	0.11	0.19	**0.20**
Blanket (15)	0.45	0.72	0.26	0.42	**0.79**	0.41	0.50	0.16	0.39	**0.77**
Bag (4)	0.54	0.65	0.31	0.53	**0.76**	0.42	0.38	0.23	0.49	**0.64**
Kanagawa (18)	0.22	0.36	0.03	0.40	**0.58**	0.05	0.15	0.01	0.38	**0.41**
Van Gogh (18)	0.29	0.67	0.04	0.46	**0.70**	0.08	**0.50**	0.01	0.45	0.46
Lascaux (18)	0.38	0.65	0.03	0.59	**0.82**	0.15	0.36	0.00	0.57	**0.76**



(a) (b)

Figure 6. Recognition rate between the reference and deformed images: (a) the Real-world blanket object sequence; (b) synthetic Kanagawa sequence. The reference images have the id 1 (on the right of each bar plot).

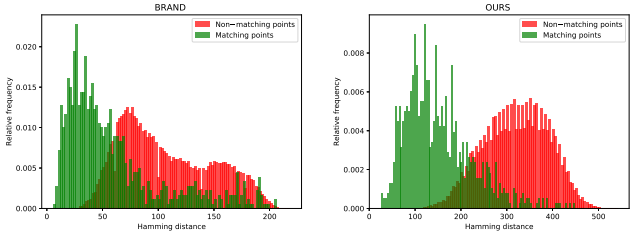


Figure 7. Histograms of Hamming distances between pairs of corresponding and non-corresponding keypoints.

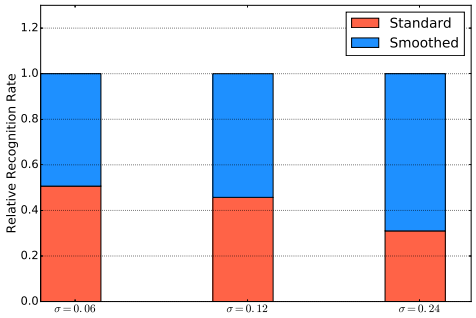


Figure 8. Relative recognition rates when using the proposed smoothing step, compared to the standard approach of directly estimating the heatflow. One can note the improvement in the recognition rate as the standard deviation of the noise increases.

tion $\mathcal{N}(0, \frac{30^2}{100})$; and ii) a uniform distribution, where we randomly selected $1,024$ different angles and isocurves.

4.4. Results

Table 1 shows the AUC and recognition rates values for all descriptors in our experiments. These experiments have shown that our descriptor is a clear winner, achieving the best performance in terms of both recognition rate and AUC. A detailed performance assessment for a real-world object and a synthetic sequence is shown in Figure 6. We note that among all methodologies, our descriptor stands out as the descriptor with the highest average in both the recognition rate and the AUC over different deformations.

We separately ran an experiment with TFEAT [29], a state-of-the-art ConvNet-based method for the local description of patches, on the real datasets. Their computa-

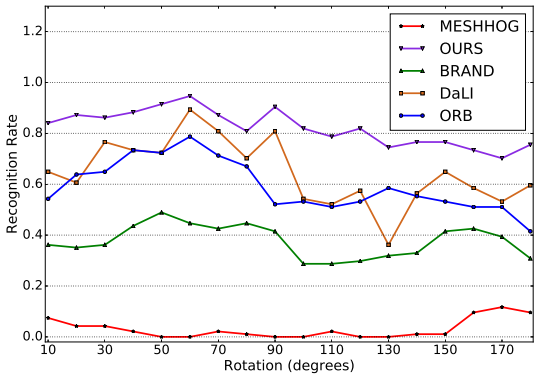


Figure 9. Recognition rate curves with respect to the rotation of each frame relative to the reference frame in the Lascaux sequence. This experiment evaluates the robustness of the descriptors to both deformation and rotation.

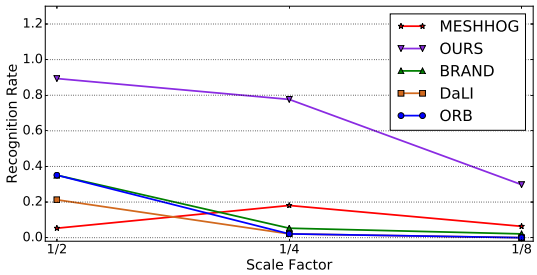


Figure 10. Demonstration of the better scale-invariance of our descriptor. The plot shows the recognition rate by a function of image scale variation for the Lascaux sequence. Even RGB-D descritors are not able to perform well in such extreme scale changes, such as 0.25 without a prior scale estimation step.

tional effort is far more expensive in terms of floating-point operations, and our method can achieve an accuracy improvement of 4.5% p.p., on average.

We can also draw the following observations. First, the poor efficiency of MeshHoG can be explained by the fact that it considers a uniformly sampled mesh to compute its descriptor, while RGB-D sensors provide noisy and non-uniformly sampled points, especially when strong deformations happen on the surfaces. Pre-processing steps can be done to mitigate this problem, however, regular area mesh decimation is generally an expensive step. BRAND performance is also reduced by deformations since its computation is based on the normals of a support region, which is not an intrinsic property of a surface, hence not being invariant to non-rigid isometric deformations. Second, the photometric information is also impaired by the deformations, which penalizes RGB-D descriptors like BRAND and MeshHOG twice.

Distance Distributions. Figure 7 shows the histograms of Hamming distances between corresponding and non-corresponding keypoints, in green and red respectively. For both descriptors, it is expected that the distribution of non-matching keypoints to be roughly represented by a Gaussian centered around the middle of the X axis. An ideal descriptor would be able to separate corresponding and non-corresponding keypoints using a threshold in the Hamming distance. In the case of overlapping between the distributions, any threshold value will lead to false positives or negatives. One can clearly see in Figure 7 that the histogram of our descriptor presents a smaller overlapping area between the distributions.

Table 2. Timing in seconds of each step of the descriptor for **94** keypoints – Intel (R) Core (TM) i7-7700 CPU @ 3.60GHz.

Method	Non-rigid	Isocurve	Extraction	Matching	Total
ORB	✗	–	0.01	0.001	0.011
BRAND	✗	–	0.31	0.001	0.311
MeshHoG	✗	–	28.52	0.030	28.550
DaLI	✓	–	61.19	6.330	67.520
Ours	✓	4.09	10.81	0.023	14.923

Robustness to noise. Figure 8 shows the relative recognition rate achieved when using the pyramid smoothing step. We tested three different levels of noise applied to the Kanagawa sequence, which contains 18 image pairs with up to 100 matching keypoints. Although we might lose fine-grained details when applying our multi-resolution strategy, the evidence of increasingly gains in recognition rates, shown in this experiment, demonstrates that the geometry of the manifold is sufficiently preserved to provide reliable geodesic distances while removing high-frequency noise, typically present in RGB-D data.

Rotation and Scale Invariance. We also pit our descriptor against other methods in terms of robustness to rotation and scale transformations. For these tests, we used the Lascaux dataset, where the camera suffers in-plane rotations ranging from 0° to 180° degrees, using a step size of 10° degrees for rotation and we applied downscale of $1/2$, $1/4$ and $1/8$. The recognition rate curve for rotation and scale transforms are shown in Figures 9 and 10, respectively. The results are given by the percentage of true matches as a function of the rotation angle and scale. It is worth noting that our descriptor outperforms all methods in all frames in both rotation and scale evaluations.

Processing Time. Table 2 shows the computation time of each step for the compared descriptors. Our method was in average 4.5 times faster than DaLI, which shows the state-of-the-art performance in matching regarding the description of deformable objects.

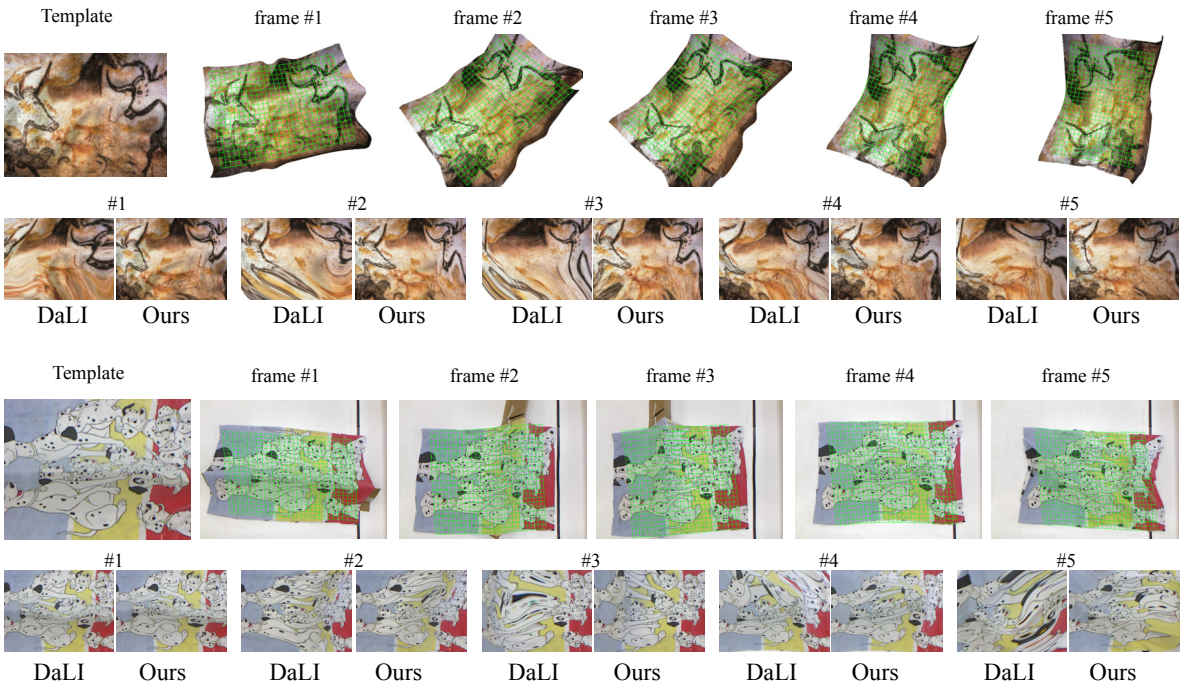


Figure 11. Tracking results of Lascaux (synthetic sequence) and Blanket (real-world sequence) using DaLI and our descriptor. The tracked region is highlighted by the green grid.

4.5. Deformable Surface Tracking Application

In this section, we evaluate the performance of our descriptor in tracking a region-of-interest of different textured meshes, subjected to large rotations, scale changes, and strong non-rigid deformations. The selection of keypoints' set $\mathcal{K}$ was done either manually or using Harris corner detector [5]. For each descriptor, we computed the Hamming distance matrix of all visible keypoints, within this region-of-interest, and performed the correspondences using the SIFT matching strategy [14], *i.e.*, the keypoint is a valid match if the ratio between the two best match candidates is smaller than a threshold (in our application we selected 0.7). Then the registration, between the template and current images during the tracking, was performed using the Deformable-Affine Thin-Plate Spline warp [1], as shown in Figure 11 for the Lascaux and Blanket sequences, using DaLI and our descriptor.

We can observe that our tracking presents better appearance quality and consistency. We also note that our binary descriptor is robust to illumination changes induced by the deformation (surface not respecting the Lambertian hypothesis) and by small specular reflexions. Please see our supplementary material pdf document and our demo video for more details and checking several full sequence tracking of different objects (paintings, shirts, and bags).

5. Conclusions

In this paper, we present GEOBIT, a novel descriptor invariant to isometric deformations, rotation, scale, and with competitive memory consumption and matching time when compared to other descriptors. Our approach combines both photometric and geometric information from RGB-D images to provide discriminative features even in the presence of non-rigid transformations. A comparative analysis was conducted against four standard descriptors and the experimental results showed that using an isometric invariant property of a manifold can be useful to create a descriptor with better matching correspondence performance. With the strategy of combining different cues, our descriptor exhibited favorable performance in matching experiments, as well as in the invariance to rotation and scale tests.

Our results extend the conclusions of [17, 11, 27], where the combined use of intensity and shape information is advantageous in improving the quality of keypoint matching. Also, the encouraging results in tracking deformable objects and high score results on the recognition rate demonstrates that the features extracted by our descriptor would be useful to improve the accuracy of the classification and recognition tasks of deformable objects.

Acknowledgments. The authors would like to thank CAPES (#88881.120236/2016-01), CNPq, FAPEMIG, and Petrobras for funding different parts of this work.

References

[1] Adrien Bartoli, Mathieu Perriollat, and Sylvie Chambon. Generalized thin-plate spline warps. *International Journal of Computer Vision*, 88(1):85–110, 2010. 8

[2] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary Robust Independent Elementary Features. In *Proc. ECCV*, September 2010. 1, 2, 3, 4, 5

[3] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Trans. Graph. (TOG)*, 32(5):152:1–152:11, Oct. 2013. 3

[4] Hao Guan and William AP Smith. BRISKS: Binary features for spherical images on a geodesic grid. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[5] Christopher Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, pages 23.1–23.6, 1988. 8

[6] Andrew E. Johnson and Martial Hebert. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. *IEEE Trans. PAMI*, pages 433–449, 1999. 1, 2

[7] Asako Kanezaki, Zoltan-Csaba Marton, Dejan Pangercic, Tatsuya Harada, Yasuo Kuniyoshi, and Michael Beetz. Voxelized Shape and Color Histograms for RGB-D. In *IROS Workshop on Active Semantic Perception*, September 2011. 2

[8] Yan Ke and Rahul Sukthankar. PCA-SIFT: A More distinctive Representation for Local Image Descriptors. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004. 5

[9] Iasonas Kokkinos, Michael M. Bronstein, Roee Litman, and Alex M. Bronstein. Intrinsic shape context descriptors for deformable shapes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 159–166, June 2012. 2

[10] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *Proc. ICRA*, May 2011. 2

[11] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Sparse distance learning for object recognition combining RGB and depth information. In *Proc. ICRA*, 2011. 2, 8

[12] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary robust invariant scalable keypoints. In *Proc. ICCV*, 2011. 2

[13] Jing Li and Nigel M. Allinson. A Comprehensive review of Current Local Features for Computer Vision. *Neurocomputing*, 71(10-12):1771–1787, 2008. 1

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004. 1, 2, 8

[15] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evalution of Local Descriptors. *IEEE Trans. PAMI*, 27(10):1615–1630, 2005. 5

[16] Francesc Moreno-Noguer. Deformation and illumination invariant feature point descriptor. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1593–1600, 2011. 2, 5

[17] Erickson R. Nascimento, Gabriel L. Oliveira, Mario F. M. Campos, Antônio W. Vieira, and William Robson Schwartz. BRAND: A Robust Appearance and Depth Descriptor for RGB-D Images. In *Proc. IROS*, 2012. 2, 3, 5, 8

[18] Erickson R. Nascimento, Gabriel L. Oliveira, Antônio W. Vieira, and Mario F. M. Campos. On the development of a robust, fast and lightweight keypoint descriptor. *Neurocomputing*, 120, 2013. 2

[19] Erickson R. Nascimento, William Robson Schwartz, and Mario F. M. Campos. EDVD - Enhanced Descriptor for Visual and Depth Data. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2012. 2

[20] Timo Ojala, Matti Pietikinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996. 1, 2, 3

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: an efficient alternative to SIFT or SURF. In *Proc. ICCV*, Barcelona, 2011. 1, 2, 3, 5

[22] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *Proc. ICRA*, pages 1848–1853, 2009. 1

[23] Gil Shamai and Ron Kimmel. Geodesic distance descriptors. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3624–3632, July 2017. 2

[24] Edgar Simo-Serra, Carme Torras, and Francesc Moreno-Noguer. DaLI: deformation and light invariant descriptor. *International journal of computer vision*, 115(2), 2015. 2

[25] Vitaly Surazhsky, Tatiana Surazhsky, Danil Kirsanov, Steven J Gortler, and Hugues Hoppe. Fast exact and approximate geodesics on meshes. In *ACM Trans. Graph. (TOG)*, volume 24, pages 553–560. Acm, 2005. 3

[26] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique Signatures of Histograms for Local Surface Description. In *Proc. ECCV*, 2010. 2

[27] Federico Tombari, Samuele Salti, and Luigi Di Stefano. A combined texture-shape descriptor for enhanced 3D feature matching. In *Proc. ICIP*, 2011. 2, 8

[28] Levi O. Vasconcelos, Erickson R. Nascimento, and Mario F. M. Campos. KVD: Scale invariant keypoints by combining visual and depth data. *Pattern Recognition Letters*, 86:83 – 89, 2017. 1

[29] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. 6

[30] Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu P. Horaud. Surface Feature Detection and Description with Applications to Mesh Matching. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Miami Beach, Florida, June 2009. 1, 2, 5